Power and limitations of conformal martingales

Vladimir Vovk

практические выводы
теории вероятностей
могут быть обоснованы
в качестве следствий
гипотез о *предельной*
при данных ограничениях
сложности изучаемых явлений

On-line Compression Modelling Project (New Series)

Working Paper #24

First posted June 2, 2019. Last revised June 24, 2019.

Project web site:
http://alrw.net

Abstract

This paper, accompanying my poster at ISIPTA 2019 (5 July 2019), poses the problem of investigating the power and limitations of conformal martingales as a means of detecting deviations from randomness. It starts from a brief review of conformal change detection, including CUSUM and Shiryaev–Roberts versions, and establishing simple validity results limiting the frequency of false alarms. It then gives several preliminary results in the direction of efficiency and discusses connections between randomness, exchangeability, and conformal martingales.

Contents

1 Introduction

A standard assumption in machine learning has been the assumption that the data are generated in the IID fashion, i.e., independently from the same distribution. This assumption is also known as the assumption of randomness (see, e.g., [11, Section 7.1] and [27]). In this paper we are interested in testing this assumption.

Conformal martingales are constructed on top of conventional machine-learning algorithms and have been used as a means of detecting deviations from randomness both in theoretical work (see, e.g., [27, Section 7.1], [4], [3]) and in practice (in the framework of the Microsoft Azure module on time series anomaly detection [28]). They provide an online measure of the amount of evidence found against the hypothesis of randomness and can be said to perform conformal change detection: if the assumption of randomness is satisfied, a fixed nonnegative conformal martingale with a positive initial value is not expected to increase its initial value manyfold; on the other hand, if the hypothesis of randomness is violated, a properly designed nonnegative conformal martingale with a positive initial value can be expected to increase its value substantially. Correspondingly, we have two desiderata for such a martingale S:

- **Validity** is satisfied automatically: S is not expected to ever increase its initial value by much, under the hypothesis of randomness.
- But we also want to have **efficiency**, i.e., we want to have S_n/S_0 large with a high probability, if the hypothesis of randomness is violated.

In the language of statistical hypothesis testing, validity corresponds to controlling the error of the first kind, and efficiency corresponds to controlling the error of the second kind (see, e.g., [12]).

Conformal martingales are defined and their validity is established in Section 2. Efficiency is not guaranteed and depends on the quality of the underlying machine-learning algorithm. It is often argued that the kind of validity enjoyed by nonnegative martingales is too strong, and we should instead be looking for a procedure of detecting deviations from randomness that is valid only in the sense of not raising false alarms too often and is efficient in the sense of raising an alarm soon after randomness becomes violated; both properties can be required to hold with high probability or on average. In the second half of Section 2 conformal martingales are adapted to such less demanding requirements of validity using the standard CUSUM and Shiryaev–Roberts procedures.

Sections 3 and 4 deal with the much more difficult question of efficiency. We ask how much we can potentially lose when using conformal martingales as compared with unrestricted testing either IID or exchangeability. This question is formalized using Cournot's principle. We will see that at a crude scale customary in the algorithmic theory of randomness we do not lose much when restricting our attention to testing randomness with conformal martingales.

Connections with the algorithmic theory of randomness will be explained in Appendix A. The main part of this paper will not use the algorithmic notion of randomness; however, as customary in the algorithmic theory of randomness, in our discussions of efficiency we will concentrate on the binary case and on the case of a finite time horizon N. These restrictions go back to Kolmogorov (cf. [26, Appendix A]); it would be interesting to eliminate them after a complete exploration of the binary and finite-horizon case (but we are not at that stage as yet).

2 Conformal martingales

First I give some basic definitions of conformal prediction (see, e.g., [27] or [1] for further details). Let $\mathbf{Z}$ be a measurable space, the space of observations.

A *nonconformity measure* is a measurable function A mapping any finite sequence $(z_1, \dots, z_n) \in \mathbf{Z}^n$ of observations of any length $n \in \mathbb{N} := \{1, 2, \dots\}$ to a sequence of numbers $(\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$ of the same length that is *equivariant* in the following sense: for any $n \in \mathbb{N}$ and any permutation $\pi : \{1, \dots, n\} \to \{1, \dots, n\}$,

$$A(z_1, \dots, z_n) = (\alpha_1, \dots, \alpha_n) \Longrightarrow A(z_{\pi(1)}, \dots, z_{\pi(n)}) = (\alpha_{\pi(1)}, \dots, \alpha_{\pi(n)}).$$

Intuitively, α_i (the *nonconformity score* of z_i) tells us how strange z_i looks as an element of the sequence $(z_1, \dots, z_n)$.

Any conventional machine-learning algorithm can be turned (usually in more than one way) into a nonconformity measure. For example, suppose each observation z_i consists of two components, $z_i = (x_i, y_i)$, where $x_i \in [-1, 1]^{16 \times 16}$ is a hand-written digit (a 16×16 matrix of pixels, each pixel represented by its brightness on the scale $[-1, 1]$) and $y_i \in \{0, \dots, 9\}$ is its label (the true digit represented by the image). The 1-Nearest Neighbour algorithm can be turned into the nonconformity measure

$$\alpha_i := \frac{\min_{j \in \{1,\dots,n\} : y_j = y_i, j \neq i} d(x_i, x_j)}{\min_{j \in \{1,\dots,n\} : y_j \neq y_i} d(x_i, x_j)}, \tag{1}$$

where $d(x_i, x_j)$ is the Euclidean distance between x_i and x_j (although the tangent metric typically produces much better results). See, e.g., [27, 1] for numerous other examples.

Let us fix a nonconformity measure A. The *p-value* p_n generated by A after being fed with a binary sequence $\omega = (z_1, \dots, z_n) \in \mathbf{Z}^*$ is

$$p_n = p_n(\omega, \theta_n) := \frac{|\{i : \alpha_i > \alpha_n\}| + \theta_n |\{i : \alpha_i = \alpha_n\}|}{n} \tag{2}$$

where i ranges over $\{1, \dots, n\}$, $\alpha_1, \dots, \alpha_n$ are the nonconformity scores for $z_1, \dots, z_n$, and θ_n is a random number distributed uniformly on the interval $[0, 1]$. The following proposition gives the standard property of validity for conformal prediction (for a proof, see, e.g., [27, Proposition 2.8]).

Proposition 1. *Suppose the observations $z_1, z_2, \dots$ are IID, $\theta_1, \theta_2, \dots$ are IID and distributed uniformly on $[0,1]$, and the sequences $z_1, z_2, \dots$ and $\theta_1, \theta_2, \dots$ are independent. Then the p-values $p_1, p_2, \dots$ as defined in (2) are IID and distributed uniformly on $[0,1]$.*

The next definition is a modification of the definition of "betting functions" in [3]. A *betting martingale* is a measurable function $F : [0,1]^* \to [0,\infty]$ such that, for each sequence $(u_1, \dots, u_{n-1}) \in [0,1]^{n-1}$, $n \geq 1$, we have

$$\int_0^1 F(u_1, \dots, u_{n-1}, u) \, \mathrm{d}u = F(u_1, \dots, u_{n-1});$$

notice that betting martingales are required to be nonnegative. A *nonnegative conformal martingale* is any sequence of functions $S_n : (\mathbf{Z} \times [0,1])^\infty \to [0,\infty]$, $n = 0, 1, \dots$, such that, for some nonconformity measure A and betting martingale F, for all $m \in \{0, 1, \dots\}$, $(z_1, z_2, \dots) \in \mathbf{Z}^\infty$, and $(\theta_1, \theta_2, \dots) \in [0,1]^\infty$,

$$S_m(z_1, \theta_1, z_2, \theta_2, \dots) = F(p_1, \dots, p_m),$$

where p_n, $n \in \mathbb{N}$, is the p-value computed by (2) from the nonconformity measure A, the observations $z_1, z_2, \dots$, and the nth element θ_n of the sequence $(\theta_1, \theta_2, \dots)$.

SPRT-type change detection

Our main concern in this section is applications of conformal prediction to change detection. A typical example of change detection is where we observe attacks, which we assume to be IID, on a computer system. When a new kind of attacks appears, the process ceases to be IID, and we would like to raise an alarm soon afterwards.

There is vast literature on change detection; see, e.g., [20] for a review. However, the standard case is where the pre-change and post-change distributions are known, and the only unknown is the time of change. Generalizations of this picture usually stay fairly close to it (see, e.g., [20, Section 7.3]). Conformal change detection relaxes the standard assumptions radically. It should be said, however, that a basic and approximate version of conformal change detection has been known since 1990: see [16].

We only consider nonnegative conformal martingales S with $S_0 \in (0,\infty)$. Ville's inequality says that, for any $c > 1$,

$$\mathbb{P}(\exists n : S_n / S_0 \geq c) \leq 1/c$$

under any IID distribution. This means that if we raise an alarm when S_n/S_0 reaches threshold c, we will be wrong with probability at most $1/c$.

We can also interpret S_n/S_0 directly as the amount of evidence detected against the first n observations being IID.

As an example, for the well-known USPS dataset of handwritten digits (see, e.g., [27, Section B.1]), the performance of a nonnegative conformal martingale

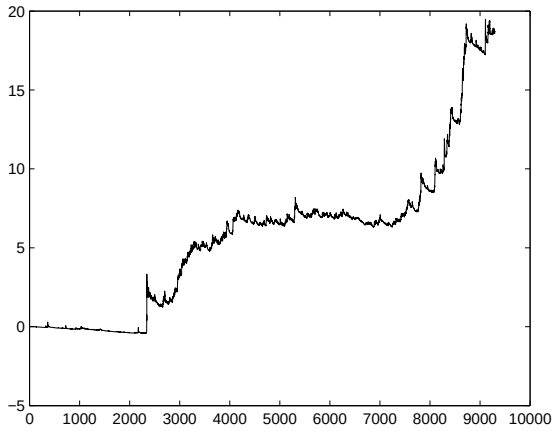

Figure 1: The values S_n of a nonnegative conformal martingale after observing the first n digits, $n = 0, \dots, 8298$, of the USPS data set, with the log-10 scale for the vertical axis. The initial value S_0 is 1, and the final value S_{8298} is 4.71×10^{18}.

based on the nonconformity measure (1) (with Euclidean metric) is shown in Figure 1 (which is Figure 7.6 in [27], where full details of the conformal martingale can be found). It is well known that the USPS data set is not random, and the lack of randomness (approximately after the 2500th observation) is detected by this conformal martingale.

CUSUM-type change detection

The kind of guarantees provided by the policy of raising an alarm when $S_n/S_0 \geq c$ is the same as that of Wald's SPRT (sequential probability ratio test), and it is often regarded as too strong to be really useful. This can be illustrated using the analogue of Figure 1 for a randomly permuted USPS dataset. The same conformal martingale performs as shown in Figure 2 (this is Figure 7.8 in [27]). The conformal martingale is trying to gamble against an exchangeable sequence of observations, which is futile, and so its value decreases exponentially quickly. If a change occurs at some point in distant future, it might take a long time for the martingale to recover its value.

A standard solution is to use the CUSUM procedure proposed by Page [18] (see also [20, Section 6.2]). According to this procedure, as applied in our current context, we raise the kth alarm at the time

$$\tau_k := \min \left\{ n > \tau_{k-1} : \max_{i=\tau_{k-1},\dots,n-1} \frac{S_n}{S_i} \geq c \right\}, \quad k \in \mathbb{N}, \tag{3}$$

where the threshold $c > 1$ is a parameter of the algorithm, $\tau_0 := 0$, and $\min \emptyset := \infty$. If $\tau_k = \infty$ for some k, an alarm is raised only finitely often; otherwise

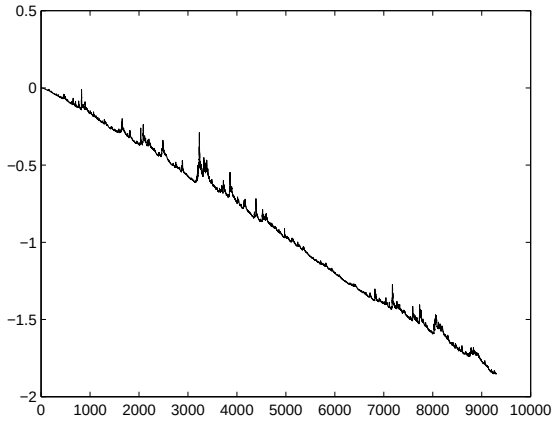

Figure 2: The values S_n of the same nonnegative conformal martingale after observing the first n digits of the USPS data set, with the log-10 scale for the vertical axis. The initial value S_0 is 1, and the final value S_{8298} is 0.0142.

it is raised infinitely often. The conformal martingale S is now additionally assumed to be positive, which ensures that the denominator in (3) is always non-zero. CUSUM is often interpreted as a repeated SPRT [18, Section 4.2]. The conformal CUSUM procedure was introduced in [25].

The following proposition gives an asymptotic property of validity of the conformal CUSUM procedure.

Proposition 2. *Let*

$$A_n := \max\{k : \tau_k \le n\}$$

be the number of alarms raised by the conformal CUSUM procedure (3) *after seeing the observations* $z_1, \dots, z_n$*. Under the assumptions of Proposition 1,*

$$\limsup_{n\to\infty} \frac{A_n}{n} \le \frac{1}{c} \quad a.s. \tag{4}$$

Under the assumptions of Proposition 1, all alarms are false, and so (4) limits the frequency of false alarms.

Proof of Proposition 2. The proof is an easy modification of the proof of the first statement of [13, Theorem 2]. Fix the threshold $c > 1$. Let $L \in \mathbb{N}$ be a large natural number (later we will let $L \to \infty$).

Set, for $n \in \{0, 1, \dots\}$,

$$\xi_n := \begin{cases} 1 & \text{if } S_i/S_n \ge c \text{ for some } i \in \{n+1, \dots, n+L\} \\ 0 & \text{otherwise.} \end{cases}$$

Let us first check that

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \xi_i \le \frac{1}{c} \quad \text{a.s.} \tag{5}$$

It suffices to check that, for each $k \in \{0, \dots, L-1\}$,

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} \xi_{k+jL} \le \frac{1}{c} \quad \text{a.s.} \tag{6}$$

Let us fix such a k. For any $j \in \{0, 1, \dots\}$, by Ville's inequality applied to the martingale S_m, $m \ge k + jL$, the probability of $\xi_{k+jL} = 1$ does not exceed $1/c$ given the values S_m, $m < k + jL$; in particular, given the values $\xi_{k+j'L}$ for $j' < j$. Now the strong law of large numbers for bounded martingale differences shows that (6) indeed holds.

It remains to deduce (4) given (5). Let us modify the definition (3) by forcing a new alarm L steps after the last alarm,

$$\tau'_k := (\tau'_{k-1} + L) \wedge \min \left\{ n > \tau'_{k-1} : \max_{i=\tau'_{k-1},\dots,n-1} \frac{S_n}{S_i} \ge c \right\}, \quad k \in \mathbb{N}, \tag{7}$$

and let

$$A'_n := \max\{k : \tau'_k \le n\}$$

be the number of alarms raised by time n. This modification can only increase A_n, $A'_n \ge A_n$, and so we have:

$$\begin{aligned} \limsup_{n\to\infty} \frac{A_n}{n} &\le \limsup_{n\to\infty} \frac{A'_n}{n} \\ &\le \limsup_{n\to\infty} \frac{1}{n} \left(\sum_{i=0}^{n-1} \xi_i + |\{i \in \mathbb{N} : \tau_i \le n, \tau_i - \tau_{i-1} = L\}| \right) \le \frac{1}{c} + \frac{1}{L} \quad \text{a.s.}, \end{aligned}$$

where the last inequality uses (5). Now let $L \to \infty$. □

Shiryaev–Roberts change detection

A popular alternative to the CUSUM procedure is the Shiryaev–Roberts procedure [23, 22], which modifies (3) as follows:

$$\tau_k := \min \left\{ n > \tau_{k-1} : \sum_{i=\tau_{k-1}}^{n-1} \frac{S_n}{S_i} \ge c \right\}, \quad k \in \mathbb{N} \tag{8}$$

(i.e., we just replace the max in (3) by $\sum$). The conformal martingale S is still assumed to be positive. The procedure defining τ_1 is based on the statistics

$$R_n := \sum_{i=0}^{n-1} \frac{S_n}{S_i},$$

which admit the recursive representation

$$R_n = \frac{S_n}{S_{n-1}} (R_{n-1} + 1), \quad n \in \mathbb{N}, \tag{9}$$

with $R_0 := 0$. An interesting finance-theoretic interpretation of this representation is that R_n is the value at time n of a portfolio that starts from \$0 at time 0 and invests \$1 into the martingale S at each time $i = 1, 2, \ldots$ [2, Section 2]. If and when an alarm is raised at time n, we apply the same procedure to the remaining observations $z_{n+1}, z_{n+2}, \ldots$.

The following proposition gives a non-asymptotic property of validity of the Shiryaev–Roberts procedure.

Proposition 3. *The conformal Shiryaev–Roberts procedure* (8) *satisfies* $\mathbb{E}(\tau_1) \geq c$ *under any IID distribution.*

Of course, we can apply Proposition 3 to other alarm times as well obtaining $\mathbb{E}(\tau_k - \tau_{k-1}) \geq c$ for all $k \in \mathbb{N}$.

Proof. The proof will follow from the fact that $R_n - n$ is a martingale; this fact (noticed, in a slightly different context, in [19, Theorem 1]) follows from (9): since S is a martingale,

$$\mathbb{E}(R_n \mid S_1, \ldots, S_{n-1}) = \frac{\mathbb{E}(S_n \mid S_1, \ldots, S_{n-1})}{S_{n-1}} (R_{n-1} + 1) = R_{n-1} + 1.$$

Another condition for $R_n - n$ being a martingale requires the integrability of R_n, which follows from

$$\mathbb{E}\left(\frac{S_n}{S_i}\right) = \mathbb{E}\left(\mathbb{E}\left(\frac{S_n}{S_i} \mid S_1, \ldots, S_i\right)\right) = \mathbb{E}(1) = 1 < \infty.$$

Fix the threshold $c > 1$. By Doob's optional sampling theorem (see, e.g., [24, Chapter 7, Section 2, Theorem 1]),

$$\mathbb{E}(\tau_1) = \mathbb{E}(R_{\tau_1}) \geq c.$$

Applying this theorem, however, requires some regularity conditions, and the rest of this proof is devoted to checking technical details.

If $\tau_1 = \infty$ with a positive probability, we have $\mathbb{E}(\tau_1) = \infty \geq c$, and so we assume that $\tau_1 < \infty$ a.s. Doob's optional sampling theorem is definitely applicable to the stopping time $\tau_1 \wedge L$, where L is a positive constant (see, e.g., [24, Chapter 7, Section 2, Corollary 1]), and so the nonnegativity of R implies

$$\mathbb{E}(\tau_1) \geq \mathbb{E}(\tau_1 \wedge L) = \mathbb{E}(R_{\tau_1 \wedge L}) \geq \mathbb{E}(R_{\tau_1} 1_{\tau_1 \leq L}) \geq c\, \mathbb{P}(\tau_1 \leq L) \to c$$

as $L \to \infty$. □

Corollary 1. *The conformal CUSUM procedure* (3) *also satisfies* $\mathbb{E}(\tau_1) \geq c$ *under any IID distribution.*

Proof. Shiryaev–Roberts raises alarms more often than CUSUM does. □

3 IID probability vs exchangeability probability

Let $\Omega := \{0,1\}^N$ be the set of all binary sequences of length N, interpreted as sequences of observations. The time horizon $N \in \mathbb{N}$ can be regarded as fixed in the rest of this paper, apart from the formulas involving $O(\dots)$, which are always uniform in N.

Let B_p be the Bernoulli probability measure on $\{0,1\}$ with the probability of 1 equal to $p \in [0,1]$: $B_p(\{1\}) := p$. The *upper IID probability* of a set $E \subseteq \Omega$ is defined to be

$$\mathbb{P}^{\text{iid}}(E) := \sup_{p\in[0,1]} B_p^N(E), \tag{10}$$

and the *upper exchangeability probability* of $E \subseteq \Omega$ is defined to be

$$\mathbb{P}^{\text{exch}}(E) := \sup_P P(E), \tag{11}$$

P ranging over the exchangeable probability measures on Ω (a probability measure P on Ω is *exchangeable* if $P(\{\omega\})$ depends only on the number of 1s in ω).

Remark. The lower probabilities corresponding to (10) and (11) are $1-\mathbb{P}^{\text{iid}}(\Omega\setminus E)$ and $1-\mathbb{P}^{\text{exch}}(\Omega\setminus E)$, respectively. In this paper we never need lower probabilities.

The function $\mathbb{P}^{\text{iid}}$ can be used when testing the hypothesis of randomness: if $\mathbb{P}^{\text{iid}}(E)$ is small (say, below 5% or 1%) and the observed sequence ω is in E that is chosen in advance, we are entitled to reject the hypothesis that the observations in ω are IID. Similarly, $\mathbb{P}^{\text{exch}}$ can be used when testing the hypothesis of exchangeability.

Proposition 4. *For any $E \subseteq \Omega$,*

$$\mathbb{P}^{\text{iid}}(E) \le \mathbb{P}^{\text{exch}}(E) \le 1.5\sqrt{N}\,\mathbb{P}^{\text{iid}}(E). \tag{12}$$

Proof. The first inequality in (12) follows from each product Bernoulli probability measure on Ω being exchangeable. If E contains either the all-0 sequence $0\dots0$ or the all-1 sequence $1\dots1$, the second inequality in (12) is obvious. If E is empty, it is also obvious. Finally, if E is nonempty and contains neither sequence, we have, for some $k \in \{1,\dots,N-1\}$,

$$\mathbb{P}^{\text{exch}}(E) = \mathbb{P}^{\text{exch}}(E\cap\Omega_k) = \frac{1/\binom{N}{k}}{(k/N)^k(1-k/N)^{N-k}}\,\mathbb{P}^{\text{iid}}(E\cap\Omega_k) \tag{13}$$

$$\le \frac{k!(N-k)!N^N}{N!k^k(N-k)^{N-k}}\,\mathbb{P}^{\text{iid}}(E) \le \sqrt{2\pi}e^{1/6}\sqrt{\frac{k(N-k)}{N}}\,\mathbb{P}^{\text{iid}}(E) \tag{14}$$

$$\le (\sqrt{2\pi}e^{1/6}/2)\sqrt{N}\,\mathbb{P}^{\text{iid}}(E) \le 1.5\sqrt{N}\,\mathbb{P}^{\text{iid}}(E), \tag{15}$$

where Ω_k is the set of all sequences in Ω containing k 1s. The first equality in (13) follows from each exchangeable probability measure on Ω being a convex mixture of the uniform probability measures on Ω_k, $k = 0,\dots,N$. The second

equality in (13) follows from the maximum of $B_p(\{\omega\})$ over $p \in [0,1]$ being attained at $p = k/N$. The first inequality in (14) is equivalent to the obvious $\mathbb{P}^{\text{iid}}(E \cap \Omega_k) \le \mathbb{P}^{\text{iid}}(E)$. The second inequality in (14) follows from Stirling's formula

$$n! = \sqrt{2\pi} n^{n+1/2} e^{-n} e^{r_n}, \quad 0 < r_n < \frac{1}{12n},$$

valid for all $n \in \mathbb{N}$; see, e.g., [21], where it is also shown that $r_n > \frac{1}{12n+1}$. The first inequality in (15) follows from $\max_{p \in [0,1]} p(1-p) = 1/4$. □

Kolmogorov's [8, 9] implicit interpretation of (12) was that $\mathbb{P}^{\text{iid}}$ and $\mathbb{P}^{\text{exch}}$ are close; on the log scale we have

$$-\log \mathbb{P}^{\text{iid}}(E) = -\log \mathbb{P}^{\text{exch}}(E) + O(\log N), \tag{16}$$

whereas typical values of $-\log \mathbb{P}^{\text{iid}}(E)$ and $-\log \mathbb{P}^{\text{exch}}(E)$ have the order of magnitude N for small (but non-zero) $|E|$. See Appendix A for further details.

4 Conformal probability

In this section we will define the upper conformal probability $\mathbb{P}^{\text{conf}}$, an analogue of $\mathbb{P}^{\text{iid}}$ and $\mathbb{P}^{\text{exch}}$ for testing randomness using conformal martingales. We will define a simple version of upper conformal probability sufficient for our current purpose; there are other natural definitions. The *upper conformal probability* of $E \subseteq \Omega$ is

$$\mathbb{P}^{\text{conf}}(E) := \inf\{S_0 : \forall (z_1, \dots, z_N) \in E : S_N(z_1, \theta_1, z_2, \theta_2, \dots) \ge 1 \text{ a.s.}\}, \tag{17}$$

where S ranges over the nonnegative conformal martingales, and "a.s." refers to the uniform probability measure over $(\theta_1, \theta_2, \dots) \in [0,1]^\infty$.

The following proposition shows that $\mathbb{P}^{\text{iid}}$ and $\mathbb{P}^{\text{conf}}$ are close, in the sense similar to the closeness of $\mathbb{P}^{\text{iid}}$ and $\mathbb{P}^{\text{exch}}$ asserted in Proposition 4 (see also (16)).

Proposition 5. *For any $E \subseteq \Omega$,*

$$\mathbb{P}^{\text{iid}}(E) \le \mathbb{P}^{\text{conf}}(E) \le N \, \mathbb{P}^{\text{exch}}(E). \tag{18}$$

Proposition 5 says that, at our crude scale, lack of exchangeability can be detected using conformal martingales. Namely, given a critical region E of a very small size $\epsilon := \mathbb{P}^{\text{exch}}(E)$, we can construct a nonnegative conformal martingale with initial capital $N\epsilon$ that attains capital of 1 when E happens.

In the rest of this section we will check Proposition 5. The following lemma asserts the left inequality in (18) (but in fact its proof proves a stronger statement).

Lemma 1. *For any $E \subseteq \Omega$, $\mathbb{P}^{\text{iid}}(E) \le \mathbb{P}^{\text{conf}}(E)$.*

Proof. We will check that the statement of the lemma remains true if the right-hand side of (17) is replaced by

$$\inf\{S_0 : \forall (z_1, \dots, z_N) \in E : \mathbb{E}\, S_N(z_1, \theta_1, z_2, \theta_2, \dots) \geq 1\}, \tag{19}$$

where the $\mathbb{E}$ refers to the uniform probability measure over $(\theta_1, \theta_2, \dots) \in [0,1]^\infty$. It suffices to prove that, for each $E \subseteq \Omega$, each $p \in [0,1]$, and each nonnegative conformal martingale S such that $\mathbb{E}\, S_N \geq 1_E$, we have $B_p(E) \leq S_0$. This follows from the property of validity (see Proposition 1) of conformal martingales:

$$B_p(E) = \mathbb{E}_{B_p}(1_E) \leq \mathbb{E}_{B_p}\, \mathbb{E}\, S_N = S_0. \qquad \square$$

It remains to check the right inequality in (18).

Lemma 2. *For any $E \subseteq \Omega$,*

$$\mathbb{P}^{\mathrm{conf}}(E) \leq N\, \mathbb{P}^{\mathrm{exch}}(E). \tag{20}$$

Proof. Let us first check the second inequality in

$$\mathbb{P}^{\mathrm{iid}}(\{\omega\}) = \frac{k^k (N-k)^{N-k}}{N^N} \leq \mathbb{P}^{\mathrm{conf}}(\{\omega\}) \leq \frac{k!(N-k)!}{N!} = \mathbb{P}^{\mathrm{exch}}(\{\omega\}), \tag{21}$$

where $k \in \{0, \dots, N\}$ and $\omega \in \Omega$ contains k 1s (all other statements in (21) were established in Proposition 4 and its proof and Lemma 1; they will not be used in the rest of this proof and are given only for symmetry).

To check the second inequality in (21), let $\omega = (z_1, \dots, z_N)$ be the representation of ω as a sequence of bits. Consider the nonnegative conformal martingale S^ω obtained from the nonconformity measure $A := 1$ and a betting martingale F such that $F(\square) = 1/\binom{N}{k}$ (where $\square$ is the empty sequence) and

$$\frac{F(p_1, \dots, p_{n-1}, p_n)}{F(p_1, \dots, p_{n-1})} := \begin{cases} \frac{n}{k_n} & \text{if } p_n \leq k_n/n \text{ and } z_n = 1 \\ \frac{n}{n-k_n} & \text{if } p_n \geq k_n/n \text{ and } z_n = 0 \\ 0 & \text{otherwise,} \end{cases}$$

where $n = 1, \dots, N$ and k_n is the number of 1s in ω observed so far,

$$k_n := |\{j \in \{1, \dots, n\} : z_j = 1\}|;$$

in particular, $k_N = k$. (Intuitively, the nonnegative conformal martingale gambles recklessly on the nth observation being z_n.) If the actual sequence of observations happens to be ω, on step n the value of the martingale is multiplied, a.s., by the fraction whose numerator is n and whose denominator is the number of bits z_n observed in ω so far. The product of all these fractions over $n = 1, \dots, N$ will have $N!$ as its numerator and $k!(N-k)!$ as its denominator. This conformal martingale is almost deterministic, in the sense of not depending on θ_n provided $\theta_n \notin \{0, 1\}$, and its final value on ω is, a.s.,

$$\frac{1}{\binom{N}{k}} \frac{N!}{k!(N-k)!} = 1.$$

To generalize (20) from singletons to arbitrary $E \subseteq \Omega$, notice that a finite linear combination of conformal martingales S^ω is again a conformal martingale, since they involve the same nonconformity measure and betting martingales can be combined. Fix $E \subseteq \Omega$ and remember that Ω_k is the set of all sequences in Ω containing k 1s. Represent E as the disjoint union

$$E = \bigcup_{k=0}^{N} E_k, \quad E_k \subseteq \Omega_k,$$

and let U_k be the uniform probability measure on Ω_k. We then have

$$\begin{aligned}\mathbb{P}^{\text{conf}}(E) \le \sum_{\omega \in E} \mathbb{P}^{\text{conf}}(\{\omega\}) &= \sum_{k=0}^{N} \sum_{\omega \in E_k} \mathbb{P}^{\text{conf}}(\{\omega\}) \le \sum_{k=0}^{N} \sum_{\omega \in E_k} \mathbb{P}^{\text{exch}}(\{\omega\}) \\ &= \sum_{k=0}^{N} U_k(E_k) \le N \max_{k=0,\dots,N} U_k(E_k) = N\, \mathbb{P}^{\text{exch}}(E),\end{aligned}$$

where the last inequality holds when, e.g., E does not contain the all-0 sequence $0 \dots 0 \in \Omega$. If E does contain the all-0 sequence, it is still true that

$$\mathbb{P}^{\text{conf}}(E) \le 1 \le N = N\, \mathbb{P}^{\text{exch}}(E). \qquad \square$$

5 Conclusion

Propositions 4 and 5 say that IID, exchangeability, and conformal upper probabilities are close, but the accuracy of these statements is very low and far from meaningful in practice. The most obvious direction of further research is to obtain more accurate results (a simple example related to Proposition 4 will be given in Appendix A). It would be ideal to establish exact bounds on upper conformal probability in terms of upper IID probability and upper exchangeability probability. The most natural definition of upper conformal probability in this context would involve randomness in a more substantial way than our official definition (17) does (cf., e.g., (19)).

References

[1] Vineeth N. Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk, editors. *Conformal Prediction for Reliable Machine Learning: Theory, Adaptations, and Applications*. Elsevier, Amsterdam, 2014.

[2] Wenyu Du, Aleksey S. Polunchenko, and Grigory Sokolov. On robustness of the Shiryaev–Roberts change-point detection procedure under parameter misspecification in the post-change distribution. *Communications in Statistics—Simulation and Computation*, 46:2185–2206, 2017.

[3] Valentina Fedorova, Alex Gammerman, Ilia Nouretdinov, and Vladimir Vovk. Plug-in martingales for testing exchangeability on-line, On-line Compression Modelling project (New Series), `http://alrw.net`, Working Paper 4, April 2012. Conference version: ICML 2012.

[4] Shen-Shyang Ho and Harry Wechsler. A martingale framework for detecting changes in data streams by testing exchangeability. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:2113–2127, 2010.

[5] Andrei N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin, 1933. English translation: *Foundations of the Theory of Probability*. Chelsea, New York, 1950.

[6] Andrei N. Kolmogorov. On tables of random numbers. *Sankhya. Indian Journal of Statistics A*, 25:369–376, 1963.

[7] Andrei N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1:1–7, 1965.

[8] Andrei N. Kolmogorov. Logical basis for information theory and probability theory. *IEEE Transactions on Information Theory*, IT-14:662–664, 1968.

[9] Andrei N. Kolmogorov. Combinatorial foundations of information theory and the calculus of probabilities. *Russian Mathematical Surveys*, 38:29–40, 1983.

[10] Andrei N. Kolmogorov and Vladimir A. Uspensky. Algorithms and randomness. *Theory of Probability and Its Applications*, 32:389–412, 1987.

[11] Erich L. Lehmann. *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, San Francisco, 1975.

[12] Erich L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer, New York, third edition, 2005.

[13] Gary Lorden. Procedures for reacting to a change in distribution. *Annals of Mathematical Statistics*, 42:1897–1908, 1971.

[14] Per Martin-Löf. The definition of random sequences. *Information and Control*, 9:602–619, 1966.

[15] Per Martin-Löf. Personal communication, February 2005.

[16] David McDonald. A cusum procedure based on sequential ranks. *Naval Research Logistics*, 37:627–646, 1990.

[17] Richard von Mises. Grundlagen der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 5:52–99, 1919.

[18] Ewan S. Page. Continuous inspection schemes. *Biometrika*, 41:100–115, 1954.

[19] Moshe Pollak. Average run lengths of an optimal method of detecting a change in distribution. *Annals of Statistics*, 15:749–779, 1987.

[20] H. Vincent Poor and Olympia Hadjiliadis. *Quickest Detection*. Cambridge University Press, Cambridge, 2009.

[21] Herbert Robbins. A remark on Stirling's formula. *American Mathematical Monthly*, 62:26–29, 1955.

[22] S. W. Roberts. A comparison of some control chart procedures. *Technometrics*, 8:411–430, 1966.

[23] Albert N. Shiryaev. On optimum methods in quickest detection problems. *Theory of Probability and Its Applications*, 8:22–46, 1963.

[24] Albert N. Shiryaev. *Probability-2*. Springer, New York, third edition, 2019.

[25] Denis Volkhonskiy, Evgeny Burnaev, Ilia Nouretdinov, Alexander Gammerman, and Vladimir Vovk. Inductive conformal martingales for change-point detection. *Proceedings of Machine Learning Research*, 60:132–153, 2017. COPA 2017.

[26] Vladimir Vovk. On the concept of the Bernoulli property. *Russian Mathematical Surveys*, 41:247–248, 1986. Russian original: О понятии бернуллиевости. Another English translation with proofs: On-line Compression Modelling project (New Series), `http://alrw.net`, Working Paper 15.

[27] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

[28] Xiao Zhang, Peter Lu, Josée Martens, Gary Ericson, and Kent Sharkey. *Time Series Anomaly Detection module in Microsoft Azure*. Microsoft, Seattle, WA, May 2019. Online documentation.

A Connections with the algorithmic theory of randomness

Propositions 4 and 5 are very crude, and Section 5 sets the task of obtaining more accurate result. This appendix explains connections of Proposition 4 with the algorithmic theory of randomness and gives references to some more precise results. It will assume knowledge of some basic notions of that theory.

The notion of randomness has been at the centre of discussions of the foundations of probability for at least 100 years, since Richard von Mises's 1919 article [17]. For von Mises, random sequences (*collectives* in his terminology) were the foundation for probability theory and statistics, and other notions, such as probability, were defined in terms of collectives.

Random sequences have been eclipsed in the foundations of mathematical probability theory by measure since Kolmogorov's 1933 *Grundbegriffe* [5]. In

the 1960s Kolmogorov started revival of the interest in random sequences, believing that they are important for understanding the applications of probability theory and statistics. He mainly concentrated on binary sequences (as a simple starting point), in which context he often referred to them as *Bernoulli sequences*. His first imperfect publication on this topic was the 1963 paper [6], but in the same year he conceived using the notion of computability for formalizing randomness. Kolmogorov's main publications on the algorithmic theory of randomness were [7, 8, 9]. He was also a co-author of [10], which was based on his ideas and publications, although he did not see the final version of that paper [10, Introduction].

Kolmogorov's notion of randomness for an element ω of a simple finite set M was that $K(\omega) \approx -\log |M|$, where K is Kolmogorov complexity and log is binary log (see [7, Section 4]). Martin-Löf [15] modified this requirement to $K(\omega \mid M) \approx -\log |M|$. In his 1968 paper [8, Section 2] Kolmogorov gave his alternative formalization of von Mises's random sequences, with a reference to Martin-Löf: namely, Kolmogorov said that a binary sequence ω of length N containing k 1s is *Bernoulli* if

$$K(\omega \mid k, N) \approx \log \binom{N}{k}.$$

It is natural to call the difference

$$d^{\mathrm{exch}}(\omega) := \log \binom{N}{k} - K(\omega \mid k, N) \tag{22}$$

the *exchangeability deficiency* of ω (in terminology close to that of [10]). Being Bernoulli in the sense of Kolmogorov does not fully reflect the intuition of being a plausible outcome of a sequence of N tosses of a possibly biased coin; this intuition is better captured by

$$d^{\mathrm{iid}}(\omega) := \inf_{p \in [0,1]} \left(-\log B_p(\omega) - K(\omega \mid p, N) \right), \tag{23}$$

which we call the *IID deficiency* of ω, being small.

Definitions (22) and (23) can be restated in terms of Martin-Löf's [14] more standard approach using nested families of critical regions. This restatement in combination with Proposition 4 immediately implies

$$d^{\mathrm{exch}}(\omega) - O(1) \le d^{\mathrm{iid}}(\omega) \le d^{\mathrm{exch}}(\omega) + \frac{1}{2} \log N + O(1). \tag{24}$$

In fact, we can interpret (24) as the algorithmic version of Proposition 4. Kolmogorov regarded the coincidence to within log as close enough, at least for some purposes: cf. the last two paragraphs of [8]; therefore, he preferred the simpler definition $d^{\mathrm{exch}}(\omega) \approx 0$ of ω being a Bernoulli sequence.

The difference between natural versions of (22) and (23) is explored in [26, Theorems 1 and 2]. Theorem 1 of [26] shows that

$$D^{\mathrm{iid}}(\omega) = \left(\log \binom{N}{k} - KP(\omega \mid N, k, D^{\mathrm{bin}}(k)) \right) + D^{\mathrm{bin}}(k) + O(1), \tag{25}$$

where k is the number of 1s in ω, KP is prefix complexity, D^{iid} is the analogue of d^{iid} using prefix instead of Kolmogorov complexity, and $D^{\mathrm{bin}}(k)$ is the *prefix binomial deficiency* of k defined by

$$D^{\mathrm{bin}}(k) := \inf_{p\in[0,1]} \left(-\log \mathrm{bin}_p(k) - KP(k \mid p, N)\right),$$

bin_p being the binomial probability distribution on $\{0,\dots,N\}$ with parameter p. Theorem 2 of [26] characterizes $D^{\mathrm{bin}}(k)$ in terms of prefix complexity, showing that it can be as large as $\frac{1}{2}\log N + O(1)$. These results can be roughly summarized as: for a binary sequence ω to be IID, it needs to be exchangeable and the number k of 1s in it needs to be binomial.

It would be interesting to state Theorems 1 and 2 of [26] without using randomness deficiency, in a form close to Proposition 4. For example, the following elaboration of the first inequality in (12) is the analogue of the inequality $\geq$ in (25).

Proposition 6. *For any $E \subseteq \Omega$,*

$$\mathbb{P}^{\mathrm{iid}}(E) \leq \mathbb{P}^{\mathrm{exch}}(E)\, \mathbb{P}^{\mathrm{bin}}(+E),$$

where

$$+E := \{z_1 + \dots + z_N : (z_1,\dots,z_N) \in E\}$$

and

$$\mathbb{P}^{\mathrm{bin}}(A) := \sup_{p\in[0,1]} \mathrm{bin}_p(A), \quad A \subseteq \{0,\dots,N\}.$$

Proof. For any $p \in [0,1]$, we have, using the notation introduced in the proof of Lemma 2:

$$\begin{aligned} B_p(E) &= \sum_{k=0}^{N} B_p(E_k) = \sum_{k=0}^{N} \mathrm{bin}_p(k) U_k(E_k) \leq \left(\sum_{k:E_k\neq\emptyset} \mathrm{bin}_p(k)\right) \max_k U_k(E_k) \\ &= \mathrm{bin}_p(+E)\, \mathbb{P}^{\mathrm{exch}}(E) \leq \mathbb{P}^{\mathrm{bin}}(+E)\, \mathbb{P}^{\mathrm{exch}}(E). \end{aligned}$$

□